International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-6, April 2020

Identifying Intrusion Behaviour using Enhanced Hidden Markov Model

T Purnima, Chandu Delhipolice, K. Sarada

Abstract: ***Data Mining is a method for detecting network intrusion detection in networks. It brings ideas from variety of areas including statistics, machine learning and database processes. Decreasing price of digital networking is now economically viable for network intrusion detection. This analysis chiefly examines the system intrusion detection with machine learning and DM methods. To improve the accuracy and efficiency of SHMM, we are collecting multiple observation in SHMM that will be called as Multiple Hidden Markov Model (MHMM). It is used to improve better Detection accuracy compare with SHMM. In the standard Hidden Markov Model, we have observed three fundamental problems are Evaluation and decoding another one is learning problem. The Evaluation problem can be used for word recognition. And the Decoding problem is related to constant attention and also the segmentation. In this Proposed Research, the primary purpose is to model the sequence of observation in Network log and credit card log transactions process using Enhanced Hidden Markov Model (EHMM). And show how it can be used for intrusion detection in Network. In this procedure, an EHMM is primarily trained with the conventional manners of a intruders. If the trained EHMM does not recognize an incoming Intruder transaction with adequately high probability, it is thought to be fraudulent.***

Keywords: IDS, KDD, HMM

I. INTRODUCTION

Global digitalization resulted in colossal data creation, both structured (“relational databases, XML”) and unstructured (“text, documents, images”). In the beginning, though manual data mining procedures used, later more mining techniques are introduced. Identifying facts, information, underlying patterns, relationships within data considered as knowledge discovery in databases (KDD). Discovering high-level information/knowledge, i.e. (“non-trivial, implicit, previously unknown and potentially useful”) from low-level data is the primary purpose of KDD. It includes multidisciplinary actions that encompass storing of data, its accessing, applying algorithms, visualization of patterns from the results, etc., DM considered as a necessary tool in knowledge discovery and decision-making process.Which paid more attention from the researchers Steps involved in KDD are as shown in Figure 1.1.

Revised Manuscript Received on April 30, 2020.
***** Correspondence Author
T Purnima*: Teaching Asst, Dep of CSE, SRM University, AP, India.
Chandu Delhipolice: Assistant Professor, Priyadarshini Institute of Technology & Science. Tenali, A.P.
K. Sarada: Research Scholar, Dep of IT, VFSTR, Vadlamudi, AP

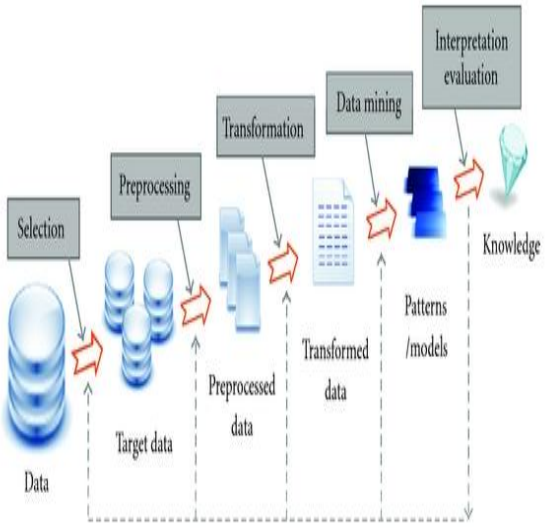

Figure 1 Overview of Steps in the KDD Process

As an initial step, the goal of the KDD process from the customer's perspective has to identify. Knowledge to be extracted as per the application domain involved needs to be understood. Target data (or a subset of data sets) as per the need are selected. This data has been cleaned using pre-processing techniques which may involve policies in handling missing fields and alter the data as per the requirements. Unwanted variables are removed and then use features analysed, which can be used to represent data that is dependent on the task. Data mining methods are applied to put forward hidden patterns and decide which models and parameters are suitable for the complete KDD process. Pattern representation, such as classification trees, regression, and clustering need to be confirmed. Essential knowledge from the mined patterns interpreted.

In this scenario, DM has been more focused on several fields such as banking and crime investigation and e-commerce and many other areas. Furthermore, it has also extended to the several other applications such as web data mining like consumer purchase tracking and stock market analysis, medical healthcare analysis, customer relations, cross-selling, product analysis, demand and supply analysis, real estates, telecommunications.

DM is a process of gaining knowledge by detailed analysis of vast sets of data. DM holds most significant potential for every industry. Most of the experts believe in DM process helps to determine risk factor, success, and failure rate in the business, which are essential factors to proceed towards our business growth.

Data Mining is an iterative process that comprises of different phases like:

1. Problem Definition: Defining the business problem
2. Data Exploration: Use of traditional data tools to explore it
3. Data Preparation: Data is tweaked multiple times in any predefined manner of modelling tools.
4. Modelling: Various modelling is applied in the data t
3. sort out the problem that needs to address.
5. Evaluation: In this, experts evaluate the data model an make changes until it satisfies the requirements.
6. Deployment: Once the evaluation implemented, minin
4. results starts.

Data Mining Tasks

5. DM tasks classified on the underlying patterns tha can mine. Data pattern according to the nature of th industry, there are two categories of functions involved i data mining.

DM errands can categorize hooked on two models, whic are descriptive and predictive.

- Descriptive
- Classification and Prediction

Descriptive model finds a relationship in data, also it functions as a means to learn more about the properties data that is examined. The predictive model determines routines on historical and existing statistics to make forecasts. Association Rules, clustering, chain discovery has been descriptive models whereas classification, regression analysis, time series analysis, forecasting and so forth are predictive models of data exploration. One of those, association rule mining technique would be your vital technique which helps people to create the ideal decisions in strengthening and improving outcome. This system calls for data and also reflects data for being a set of association rules from database centered on statistical measures distributed by this user. So, taking advantage of outsourcing data mining services and implement results in gaining more from the business is a smart move for every industry vertical like finance, legal, logistic, etc.

1.2 Data Mining Techniques

In wide-ranging explanation, DM techniques are deploying to extract information from large datasets. Furthermore, Data mining is a significant practice to ascertain knowledge about customer behavior towards business aids. It investigates regular indefinite patterns that are significant for viable achievement. However, Data mining has often taken the wrong way; people think that it comprises only handling of data but is inherently far more than this, i.e., which uses advanced tools and technologies. Furthermore, we used different techniques for different problems which depend on provided solutions. Extraction of the knowledge from massive databases, with experience. The biggest challenge is to examine the data to extract meaningful information that can be used to solve a problem or for the growth of the business.

There are a few techniques which exertion to lead the business successfully.

1. **Outlier Detection:** A primary or the foremost step in data mining which used for searches of similar items in data set to project at some pattern.
2. **Associative Technique:** To find out the relationship between variables in the large database. This technique works at a large extent, especially in large scale business for excellent data management.

Clustering Analysis: This one used when an organization wants to identify the data items which are very similar to each other. This analyzation gives the idea to find both similarities and differences.

Regression Analysis: Another tool which can be beneficial in the accounting and finance process. It works to find dependencies between database items.

Neural Networking: It includes two types. First, biological Neural, as human brains which extremely good to classify the patterns and predictions. Second, artificial which are logically programmed and implementation will do on the computer systems. In this category, the software works for quality business management.

II. BACKGROUND WORK

All things considered, ought to perform well in displaying client activities. He reasoned that the HMM and the case-based student referenced above, prepared utilizing similar Data, performed equivalently.

[2] They noticed that best execution was acquired by utilizing various states comparing to the quantity of framework calls utilized by an application.

While HMMs performed superior to anything inductive standards or succession coordinating, the creators addressed if gradual improvement merited the essentially more (days versus minutes) preparing times, and we contend that it does, particularly given the consistently expanding capacities of CPUs and the need to break down continually expanding measures of traffic. Various different methodologies have been recommended that might be valuable.

[4] drafted two interruption recognition strategies for versatile specially appointed systems, which utilizes community endeavours of hubs in an area to distinguish a malevolent hub in that area. Messages are passed between hubs and relying upon the messages got, these hubs decide suspected (hubs that are suspected to be vindictive). These speculated hubs are in the long run sent to the screen hub (the initiator of the discovery calculation).

[5] proposed a Layered IDS Framework (LIDF) to recognize bargained and noxious hubs in a specially appointed system. LIDF comprises of three modules to be specific, accumulation, identification and alarm. These modules work locally in each hub of a system. The accumulation and capacity of review information is performed with the utilization of a parallel tree. The discovery is accomplished with Lagrange inserting polynomials and the alarm is cultivated with straight edge plans.

[6] portrayed a component configuration-based model for secure pioneer decision within the sight of narrow-minded hubs. To adjust asset utilization of the hubs in system,

hubs with the most outstanding assets ought to be chosen as the pioneers. This model has presented a two head decision calculation, to be specific Cluster Dependent Leader Election (CDLE) and Cluster Independent Leader Election (CILE).

III. PROBLEM STATEMENT

The IDS is most important issue in network. The main aspires are Firstly, to identify different kinds of intrusions and secondly to estimate of unconventional techniques which are used to intrusion detection. The sub-target is to compare fraudulent detection according to recent IDS detection. Detecting IDS has made many techniques developed by a group of investigative interest groups, and special importance on Data Mining, and Machine Learning is recommended. Apart from this, there is a problem; Data possibly will be missing in in different intervals. After that, there are several observational approaches that do not essentially synchronous with each other and probably will have various “emission distributions” for same states.

In the recent year’s one of the techniques is Hidden Markov Model (HMM) has been introduced to provide a simple and active framework for modelling time-changing spectral vector sequences. One advantages of HMMs False Positives transactions identified as malicious by an FDS although they are truly genuine. Application of HMMs to speech recognition has seen extensive success and gained much reputation. However, there are various problems found in the HMM model is listed below

1. HMM is situated upon the Markov property, which situations which the reality to be at a given state at time t only depends upon their condition at time t1. This isn't always true for language noises at which dependencies some scenarios stretch through different nations.
2. Level of data Required to Coach an HMM Is Quite huge as Due to Selection of parameters to be projected within Normal collection of HMMs, high physical exercise data is difficult available.

To solve above listed problem, alternative approach for HMM is, Semi-Hidden Markov Model (SHMM)algorithm of Anomaly discovery is offered which computes the distance between processes observed by IDS and perfect normal processes. With this we are employing additional method for IDS is that having a key element is to factorize marginal log-likelihood by a variation distribution over latent variable. However, there is limitations with this approach. A major constraint is found, in Mutually HMM and SHMM i.e., it is normally imagined that there survives at least one observation associated with every state that the hidden Markov chain takes on. We distributed Data mining and Machine Learning techniques to produce classes in parallel by label transaction However, there are some limitations in SHMM are listed below

1. Amount of parameters Intrusion required to create an SHMM is tremendous.
2. Learning from mistakes way of selecting a version system topology. Findings of distinct research workers reveal that left-to-right structure implements better that ergodic. However, there isn't any formal way of determining up on structure to solving a challenge.

To improve IDS efficiency in SHMM we are using more than one observation settings, which cannot sync and inspect the semi-Hidden state layout and insight, this approach will be called as Multiple Hidden Markov Model (MHMM). Using this approach, we can improve Intrusion detection accuracy compare to SHMM. However there some drawbacks in this approach given below

1. Constant length Monitoring frames, This requirement limits chances feature extraction.
2. Parameter estimation/re-estimation problem

IV. METHODOLOGY

The main objective of this research is to introduce innovative ways to detect Intrusion behaviour in network using Data Mining and Machine Leaning.

1 To analyse Intrusion detection system with Machine Learning techniques.
2 To identify Intrusion detection system using Hidden Markov Model and its types Semi Hidden Markov Model (SHMM), Multiple- Semi Hidden Markov Model (MSHMM) and Enhanced Hidden Markov Model (EHMM).
3 To Classify intrusion through collective analysis of Network log and credit card transaction data.
4 To compare accuracy of IDS with different Hidden Markov Models.

V. INTRUSION DETECTION SYSTEM USINGSEMI-HIDDEN MARKOV MODEL(SHMM)

In this study, Health States were used to observe the possibilities of state transition prospects using modern possibilities of state states and SHMM. A modified future algorithm for SHMMs will be weighted using visual verb analysis using SHMM parameters. The prediction calculator of state-of-the-art model was provided for the rest of life. The results show that SHMM can provide valuable time information in an operator case, while team's complexity strengthens the SHMM team.

Detection algorithm

Most of Ethiopian's principles (MEP), we know that when a computer system is running in a normal system, audition data has little information, it creates it. When he runs in cheating state. For example, Intrusion information is more than the general state, so the up-to-date working card can be used as IDS. But the size visible logo layout growing logo. It only feels to entry of same length. To use the symptoms of nuclear metric variables, we determine average information of the symptoms that are considered and generally as a metric to isolate behavioural behaviour and behavioralbehavior. Considered. Use

Training Algorithms

Because the general condition of computer system may change over time, so semi-semi-semi-skull models detecting evolution. Printed Semi Markov model K = (N, M, V, A, B, π), the distribution of state transfer possibilities and the fixed state and Intrusion prices have fixed prices. ,

Then the distribution mark for BP = {b (k)}, 1 ≤ K ≤ M can be applied regularly by the administrator of the training system..

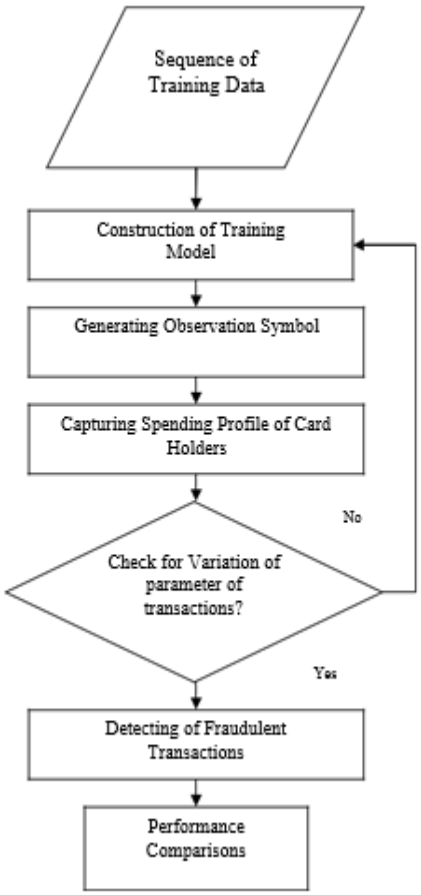

Figure 2: System Flow Diagram

ALGORITHM STEPS

TRAINING PHASE: Cluster creation

STEP 1: Identify the profile of cardholder from their purchasing

STEP 2: The probability calculation depends on the amount of time that has elapsed since entry into the current state.

STEP 3: Construct the training sequence for training model

DETECTION PHASE: Intrusion detection

STEP 1: Generate the observation symbol B

STEP 2: Form new sequence by adding B in existing sequence

STEP 3: Calculate the probability difference and test the result with training phase

STEP 4: If both are same it will be a normal customer else there will be Intrusion signal will be provided.

VI. EXPERIMENTAL RESULTS

The proposed approach performance software is estimated by using Dotnet Technology using visual studio software through credit card transaction data. Performance of the algorithm is based on the following factors

1) Precision 2) Recall 3) F-Measurement

This chart shows the F-measure Rate of Previous and Implemented work based on two parameters of F measure and number of Dataset. Using with chart we can view, when the number of Dataset is Maximized F step speed additionally Implemented in suggested work nevertheless when variety of Data Set is enhanced the F step speed is lessened in previous system than the proposed work. By this chart we can finalize that the F- measure rate of proposed work is increased which one the best of our knowledge. The values of are given in below table:

Table 1. Dataset Vs F-Measure between EHMM & SHMM

S.No	No of Datasets	EHMM	SHMM
1	10	0.90	0.78
2	20	0.84	0.74
3	30	0.79	0.65
4	40	0.67	0.55
5	50	0.56	0.45
6	60	0.43	0.34

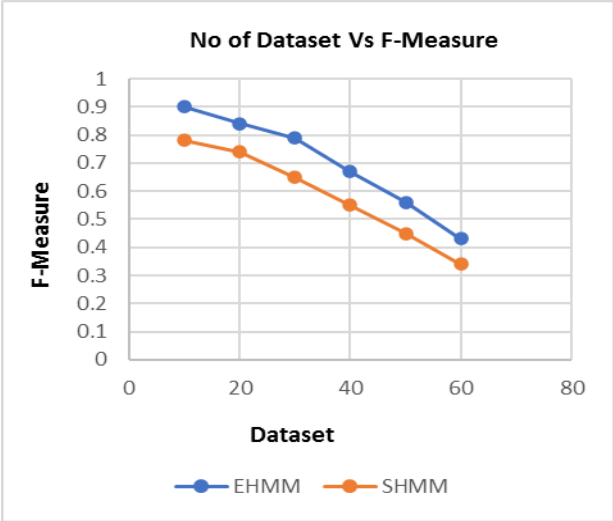

Figure 3. Dataset Vs F-Measure between EHMM & SHMM

Table 2. Data set Vs Precision between EHMM & SHMM

S. No	No of Datasets	EHMM	SHMM
1	10	0.30	0.22
2	20	0.50	0.43
3	30	0.60	0.53
4	40	0.66	0.60
5	50	0.72	0.65
6	60	0.80	0.68

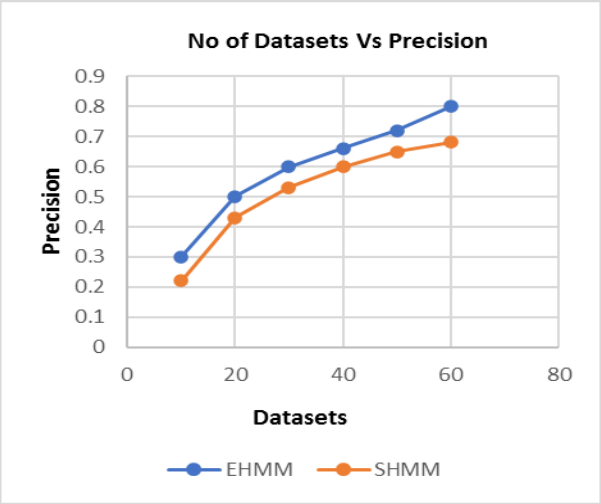

Fig 4. Data set Vs Precision between EHMM & SHMM

Table 3 Data set Vs Recall between EHMM & SHMM

S. No	No of Datasets	EHMM	SHMM
1	10	0.40	0.25
2	20	0.55	0.44
3	30	0.65	0.56
4	40	0.69	0.63
5	50	0.78	0.68
6	60	0.88	0.70

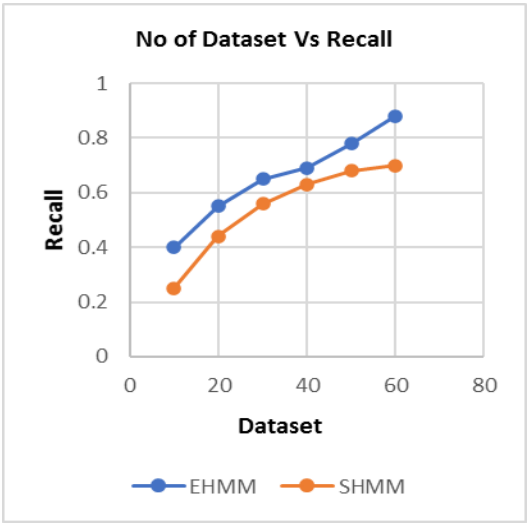

Fig 5 Data set Vs Recall between EHMM & SHMM

Table 4. Recall Comparison between EHMM & HMM

S. No	No of Datasets	EHMM	HMM
1	10	0.35	0.23
2	20	0.43	0.32
3	30	0.50	0.42
4	40	0.59	0.49
5	50	0.72	0.60
6	60	0.87	0.68

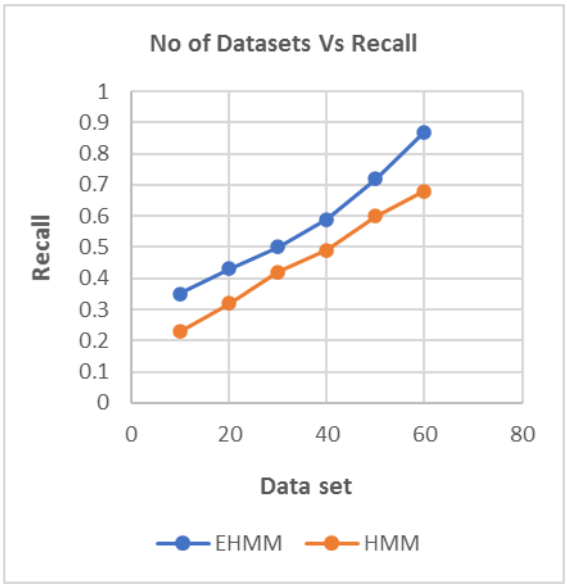

Fig 6 Recall Comparison between EHMM & HMM

VII. CONCLUSION

Recommended work wants to handle model parameters to boost your monitoring, yet one many factor presents in Standard HMM and SHMM is left-right models is that you cannot utilize one tracking deal for re-estimation of those variation parameters. It has found a very first quotation HMM parameters really are important issue and resolved. The experimental results reveal we are able to expressively reduce loss as a consequence of Intrusion through DM of all Intrusion variations during this projected endeavour of EHMM. One of those substantial disadvantages of former techniques is to calculate model parameters expressively that affects the operation detection accuracy, thus we've implemented a publication called as Improved Hidden Markov Model for maximizing model parameters. Using this specific algorithm quantity of nations and Distribution of nations chances is certain. The EHMM can reveal increased speed compared to previous Hidden Markov Models. Finally, dependent on contrast and also outcome from experimentation series suggested method works a lot better compared to other previous approaches with higher degrees and greater accuracy in detection of IDS plus it shows improved performance compared to previous work.

REFERENCES

1. A. Isazadeh,, T. Saderi, and A. H. Navin (2008), "Data-Mining by Probability-Based Patterns," pp. 353– 360.
2. B. Liu, Y. Xiao, F. Deng (2013) "SVDD-based outlier detection on uncertain data", Vol. 34, Issue 3, PP: 597–618.
3. D. Lewis, & D. Madigan (2007), "Large-Scale Bayesian Logistic Regression for Text Categorization," Vol: 49, Issue: 3, PP: 291–304.
4. Mohammed, Bhattacharya, P (2011), "Mechanism Design-Based Secure Leader Election Model for Intrusion Detection in MANET", Vol: 8, Issue: 1, PP:89-103.
5. N. Wickramasinghe, J. N. D. Gupta (2005), "Knowledge Management in Healthcare," Vol 63, PP: 5–18.
6. Otrok, Wang, L, Debbabi, M & Bhattacharya, P (2008), "A game theoretic intrusion detection model for mobile ad hoc networks", Vol 31, Issue: 4, PP: 708-721.
7. R. Armañanzas, P. Larrañaga (2013), "Unveiling relevant non-motor Parkinson's disease severity symptoms using a machine knowledge approach,". Vol: 58, Issue: 3, PP: 195– 202.

8. S. H. Liao, P. Y. Hsiao (2012), "DM methodologies and applications" Vol. 39, issue. 12, PP: 11303–11311.
9. S. W. Fei (2010), "Diagnostic study on arrhythmia cordis based on particle swarm optimization-based support vector machine," Vol. 37, Issue 10, PP: 6748–6752.
10. W.-C. Yeh (2009), "A new hybrid approach for mining breast cancer pattern using discrete particle swarm optimization and statistical method," Vol 36, Issue: 4, PP: 8204–8211.

AUTHORS PROFILE

Mrs T Purnima did her M Tech CSE in 2017 in JNTU kakinada. Total teaching experience is 3 years.Her research interests are Data Mining and machine learning.

Chandu Delhipolice Completed M Tech in JNTU Kakinada, having 6 years of teaching experience, and interested Research domains are Data Mining and Artificial Intelligence.

K Sarada Completed MCA from Acharya Nagarjuna university and M. Tech from Jawaharlal Nehru Technological University Kakinada. And presently she is research scholar in VFSTR. Her Research areas Data Mining, Machine learning, Nlp etc.